



(12) **EUROPEAN PATENT APPLICATION**

(88) Date of publication A3:
07.06.2006 Bulletin 2006/23

(51) Int Cl.:
G06F 17/27 (2006.01)

(43) Date of publication A2:
16.06.2004 Bulletin 2004/25

(21) Application number: **03028508.4**

(22) Date of filing: **10.12.2003**

(84) Designated Contracting States:
**AT BE BG CH CY CZ DE DK EE ES FI FR GB GR
HU IE IT LI LU MC NL PT RO SE SI SK TR**
Designated Extension States:
AL LT LV MK

(72) Inventors:
• **Privault, Caroline**
38700 Corenc (FR)
• **Poirier, Herve**
38700 La Tronche (FR)

(30) Priority: **13.12.2002 US 248057**
13.12.2002 US 248058

(74) Representative: **Grünecker, Kinkeldey,**
Stockmair & Schwanhäusser
Anwaltssozietät
Maximilianstrasse 58
80538 München (DE)

(71) Applicant: **Xerox Corporation**
Rochester,
New York 14644 (US)

(54) **Method and apparatus for recognizing multiword expressions**

(57) Words of an input string are morphologically analyzed to identify their alternative base forms and parts of speech. The analyzed words of the input string are used to compile the input string into a first finite-state network. The first finite-state network is matched with a second finite-state network of multiword expressions to identify all subpaths of the first finite-state network that match one or more complete paths in the second finite-

state network. Each matching subpath of the first finite-state network and path of the second finite-state network identify a multiword expression in the input string. The morphological analysis is performed without disambiguating words and without segmenting the input string into sentences in the input string to compile the first finite-state network with at least one path that identifies alternative base forms or parts of speech of a word in the input string.

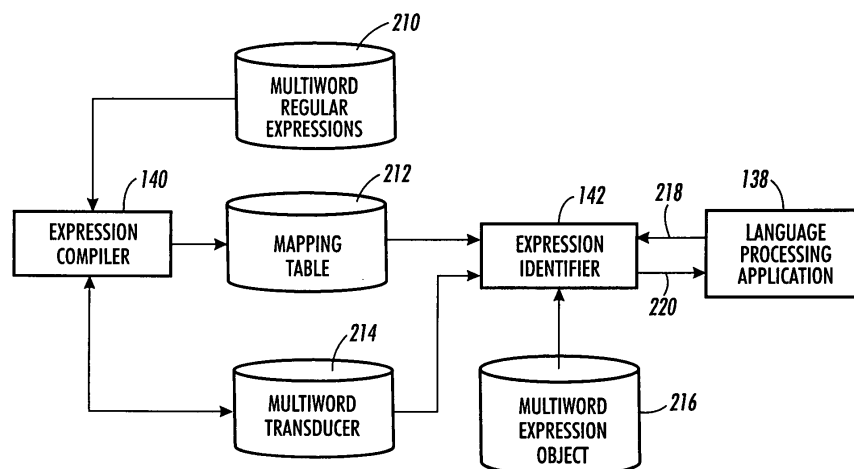


FIG. 2



European Patent
Office

EUROPEAN SEARCH REPORT

Application Number
EP 03 02 8508

DOCUMENTS CONSIDERED TO BE RELEVANT			
Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
X	JEAN-PIERRE CHANOD AND PASI TAPANAINEN: "A Lexical Interface for Finite-State Syntax" TECHNICAL REPORT MLTT-025, [Online] 9 February 1996 (1996-02-09), XP002377661 Rank Xerox Research Centre, Grenoble, France Retrieved from the Internet: URL:http://citeseer.ist.psu.edu/458124.htm 1> [retrieved on 2006-04-20]	1-4,9-11	INV. G06F17/27
Y	* the whole document *	5-8	
X	CHANOD J-P ET AL: "A Non-Deterministic Tokeniser for Finite-State Parsing" WORKSHOP ON EXTENDED FINITE STATE MODELS OF LANGUAGE. ECAI, XX, XX, 11 August 1996 (1996-08-11), pages 1-3, XP002298629 * the whole document *	1,10,11	
X	US 6 393 389 B1 (CHANOD JEAN-PIERRE ET AL) 21 May 2002 (2002-05-21) * the whole document *	1,10,11	TECHNICAL FIELDS SEARCHED (IPC) G06F G10L
D,Y	US 5 950 184 A (KARTTUNEN ET AL) 7 September 1999 (1999-09-07) * the whole document *	5-8	
D,A	ELISABETH BREIDT, FRÉDÉRIQUE SEGOND AND GUISEPPE VALETTI: "IDAREX: Formal Description of Multi-Word Lexemes with Regular Expressions" MLTT-022, [Online] November 1995 (1995-11), pages 1036-1040, XP002377662 Retrieved from the Internet: URL:http://www.cs.mu.oz.au/acl/C/C96/C96-2182.pdf> [retrieved on 2006-04-20] * the whole document *	1,9-11	
The present search report has been drawn up for all claims			
Place of search Berlin		Date of completion of the search 20 April 2006	Examiner Woods, J
CATEGORY OF CITED DOCUMENTS X : particularly relevant if taken alone Y : particularly relevant if combined with another document of the same category A : technological background O : non-written disclosure P : intermediate document		T : theory or principle underlying the invention E : earlier patent document, but published on, or after the filing date D : document cited in the application L : document cited for other reasons & : member of the same patent family, corresponding document	

3
EPO FORM 1503 03.82 (P04C01)

**ANNEX TO THE EUROPEAN SEARCH REPORT
ON EUROPEAN PATENT APPLICATION NO.**

EP 03 02 8508

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

20-04-2006

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 6393389	B1	21-05-2002	NONE

US 5950184	A	07-09-1999	DE 69615794 D1 15-11-2001
			DE 69615794 T2 18-04-2002
			EP 0775963 A2 28-05-1997
			JP 9179872 A 11-07-1997
